FairDeDup: Detecting and Mitigating Vision-Language Fairness Disparities in Semantic Dataset Deduplication

Eric Slyman Oregon State University
Stefan Lee Oregon State University
Scott Cohen Adobe
Kushal Kafle Adobe

Oregon State University · Adobe · CVPR SEATTLE, WA JUNE 17-21, 2024 · Project Page

TL;DR: Training large vision-language (VL) models is **expensive.** Our work enables training **more quickly** and with **fewer resources** by removing similar data points. Additionally, our method **mitigates bias** by preserving human-defined dimensions of diversity during this deduplication process.

Web-Scale VL Models Are Expensive

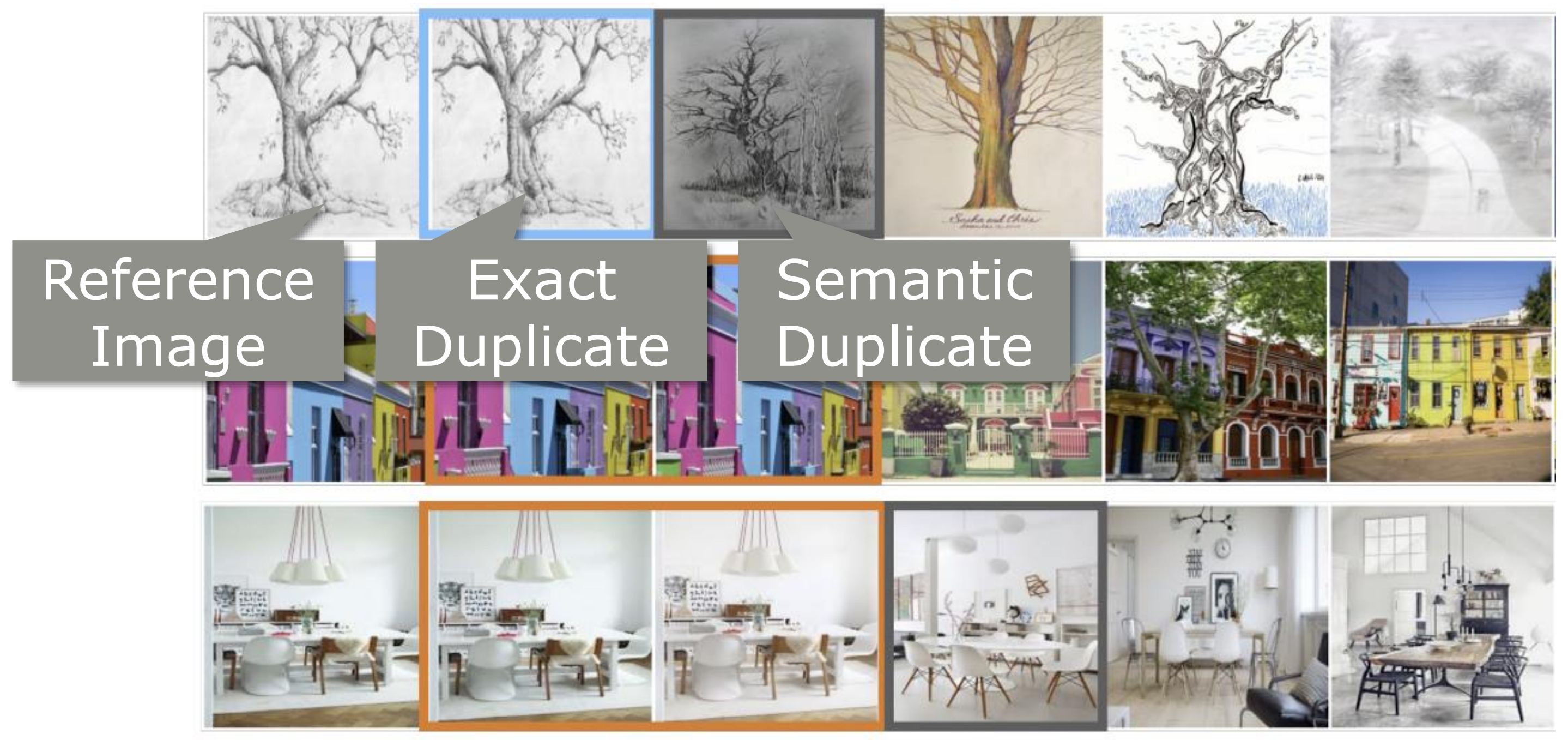
Leading web-scale VL models are trained over billions of samples and use hundreds of GPUs. LAION used **824 GPUs** for 11 days to train **one large model** (CLIP H/14-4B).

The on-demand AWS EC2 GPU cost alone for this training is

$870K over 11 Days

Web Training Data Contains Duplicates

Removing copies of the same data can improve training efficiency by preventing redundant passes over the same information. SemDeDup (SDD) [1] is a SOTA deduplication allowing **half the original amount of training cost and time** with **negligible degradation in model accuracy.**

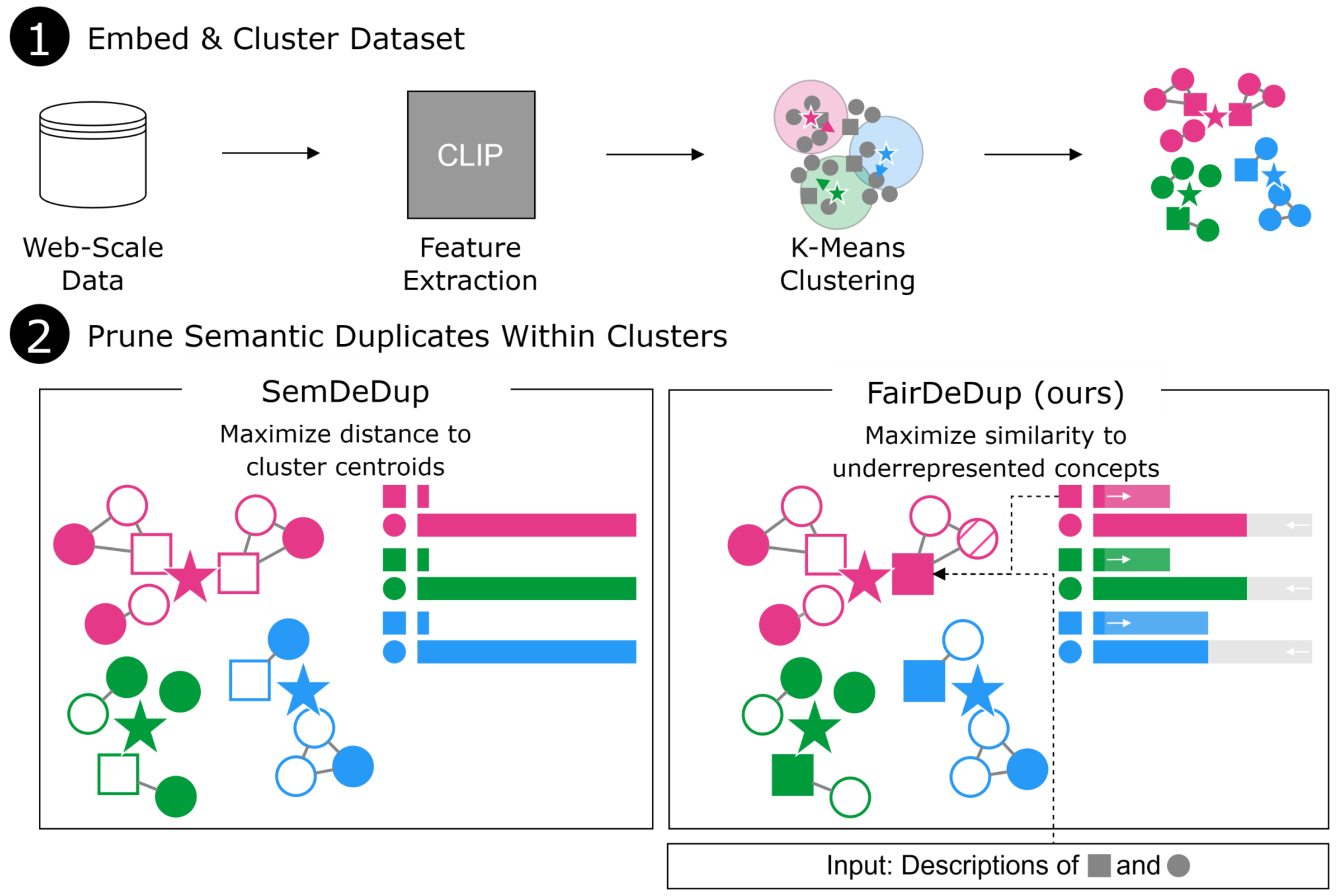


Deduplication Can Reinforce Bias

We find that semantic deduplication can exasperate fairness disparities. For example, reducing the range of genders, races, or ages seen in preserved images depicting certain occupations.

Our Solution: FairDeDup (FDD)

FDD preserves data diversity along human-defined semantic concepts. These concepts are specified in natural language and are incorporated into the heuristic used for selecting samples to preserve from neighborhoods of duplicated data.


1 Embed & Cluster Dataset

2 Prune Semantic Duplicates Within Clusters

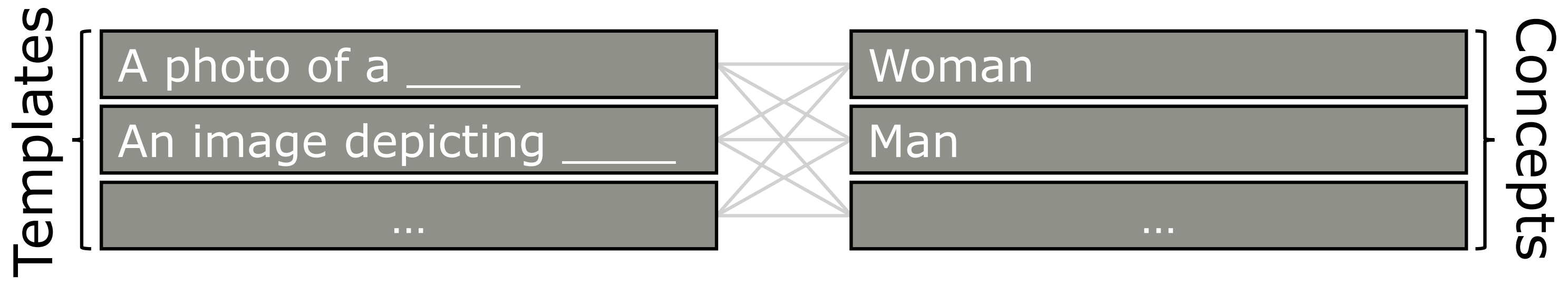
SemDeDup: Maximize distance to cluster centroids

FairDeDup (ours): Maximize similarity to underrepresented concepts

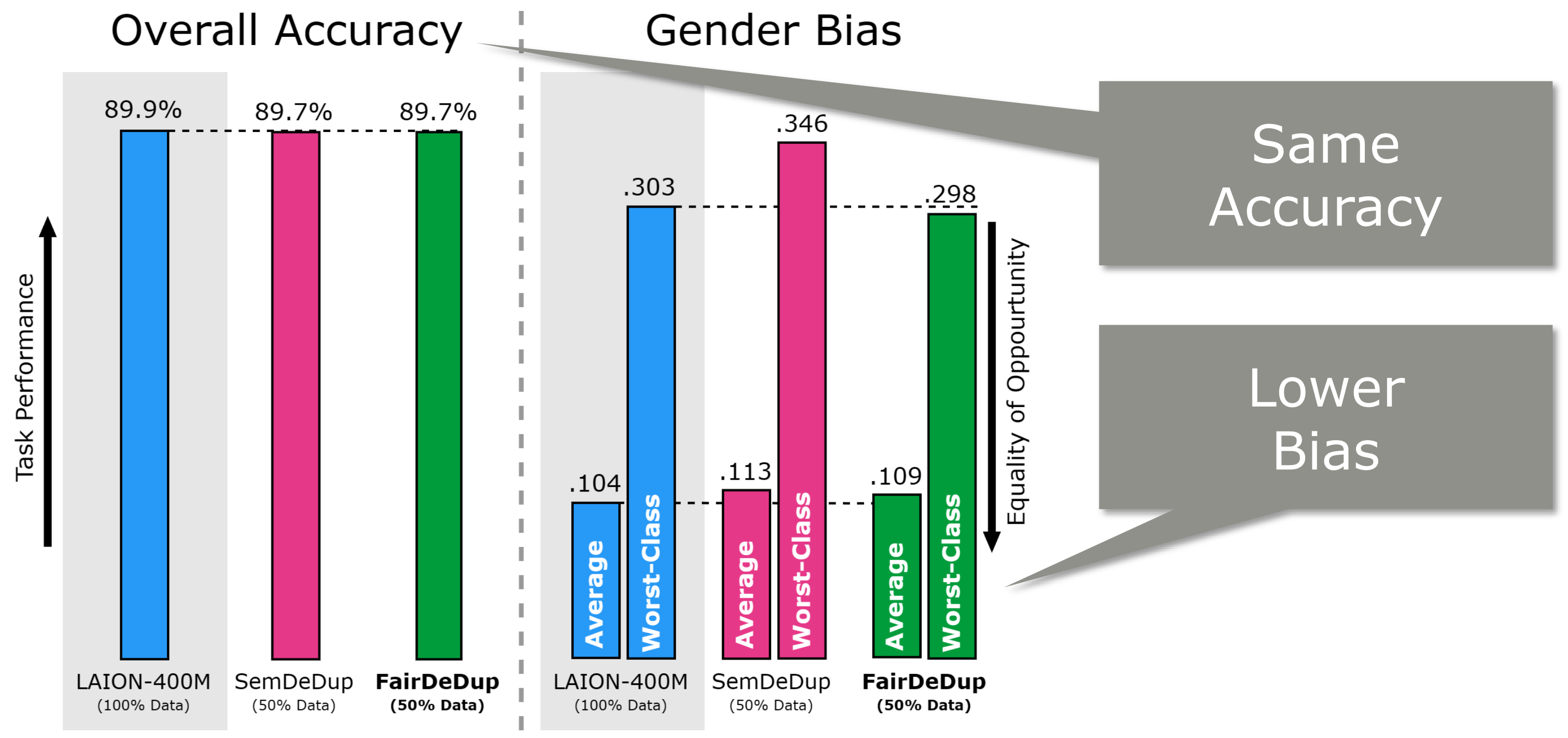
Input: Descriptions of ■ and ●

Example: Mitigating Gender Bias

A machine learning engineer may choose, for example, to write several templated captions describing different genders. FDD then uses these captions to ensure these genders are equally included in the data after pruning.


Templates: A photo of a _____ · An image depicting _____ · ...

Concepts: Woman · Man · ...



Same Accuracy

Lower Bias

FDD Preserves The Data Distribution

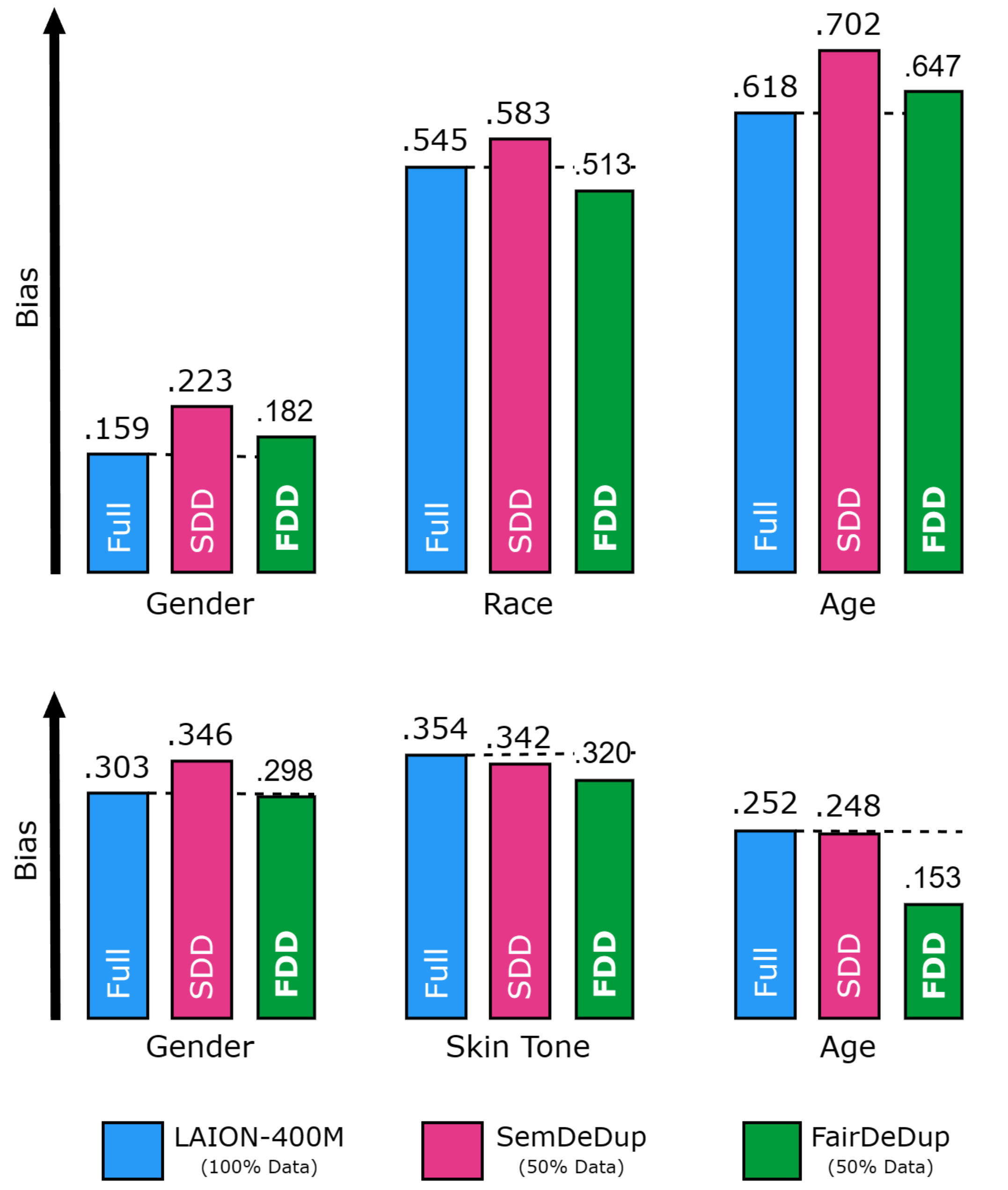
Running deduplication on data with human-annotated demographic information shows that SDD overly prunes already underrepresented groups of people. FDD more closely matches the distribution of data assigned to underrepresented groups with >99% confidence (n=10).

FDD Reduces Downstream Bias

When evaluated against SDD, we find that **FDD improves fairness outcomes** in nearly every case **without negatively impacting accuracy or requiring additional training.** The best performing deduplicated model is in **bold**. Lower is better.

FairFace:
Bias Against Most Disadvantaged Group In Text-Based Image Query

FACET:
Worst-Case Bias Against a Group In Text-Specified Image Classification



[1] Abbas et al. SemDeDup: Data-Efficient Learning at Web-Scale Through Semantic Deduplication. arXiv preprint, 2023.